

Nuclear Import and Sorting of Proteins

Hippolyt Ritter

LMU/TU München

Received on 17/06/2013

Advisor: Tatyana Goldberg

ABSTRACT

Motivation: The nucleus plays a key role in the eukaryotic cell. Therefore it is crucial to understand nuclear import and the sorting of proteins as regulatory processes that determine the nucleus' function. As experimental knowledge of both processes is incomplete, computational models are needed.

Results: This report summarises and presents the most important ideas and results from two papers. One by Mehdi et al. develops a probabilistic model of nuclear import [2] and the other one by Bauer et al. tries to predict the subnuclear compartment association of nuclear proteins[1].

1 INTRODUCTION

The cell nucleus is the defining feature of the eukaryotic cell. As host to the genome it carries the genetic information of the organism and tightly regulates all processes involving DNA including DNA replication, transcription, but also steps in RNA processing such as capping and splicing.

A first step of regulation of nuclear processes is nuclear import: The nucleus is separated from the rest of the cell by a membrane. Therefore proteins, localising in the cytoplasm after translation, have to be translocated into the nucleus. The most common and best understood pathway makes use of so called classical nuclear localisation signals (cNLS). These cNLS, short motifs in the amino acid sequence of proteins, are recognised and bound by importin- α . Importin- α in turn is bound by importin- β (sometimes these are also referred to as karyopherins). The complex of importin- α , importin- β and the cargo protein can then be recognised by the nuclear pore complex and is translocated into the nucleus. There the cargo protein is set free. The protein Ran is involved in this process.

In the nucleus proteins assemble to subnuclear compartments. These are highly dynamic (i.e. they are not present in constant number in all cells/cell types at every stage of the cell cycle) complexes of DNA, RNA and protein. The most prominent and best understood one is the nucleolus. After mitosis up to four nucleoli form in the cell. They assemble around clusters of ribosomal genes and are involved in transcribing rRNA and biogenesis of ribosomes.

2 BAYESIAN NETWORKS

As both Mehdi and Bauer use Bayesian Networks in their model they shall be introduced here briefly. Bayesian Networks are Directed Acyclic Graphs in which the nodes represent random

variables and the edges represent their dependencies (parent to child). The probability of a certain state of a network with variables $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$ can be computed with the formula

$$P(x_1, \dots, x_n) = \prod_{i=1}^N P(x_i | pa(X_i)) \quad (1)$$

with $pa(X_i)$ being the set of nodes that have a directed edge towards X_i .

Nodes can take boolean or continuous values. For nodes with boolean parents only, the probability of a certain state basically results from a conditional dependency table created in training.

Bayesian Networks can also include so called latent variables for which no values can be observed. Their values are therefore estimated using the EM algorithm that computes parameters to estimate the value of a latent variable from the state of the parents so that the overall probability of the model is optimised in training.

3 A PROBABILISTIC MODEL OF NUCLEAR IMPORT

Mehdi et al. [2] made use of the knowledge about nuclear import and created a probabilistic model to predict the import of proteins into the nucleus. They built their model as Bayesian Network which not only allows to assign a probability to nuclear import, but also provides mechanistic reasons for the prediction, i.e. it shows which parts of the model are responsible for the value of the probability.

The model was tested via cross-validation on a mouse and a yeast data set and on an independent data set. The quality of its predictions was compared to those of other state-of-the-art tools.

3.1 Model

The model (Fig. 1) consists of three main modules that make a prediction based on their available information: The cNLS-only module, which rates the presence of nuclear localisation signals for the importin- α pathway, the PPI-NLSdb module, which checks for interaction with importin- α , importin- β and Ran and for the presence of an alternative nuclear localisation signal, and the SVM-sequence module, which rates the sequence similarity with known nuclear proteins based on a support vector machine.

3.1.1 The cNLS-only module The core of the cNLS-only module are four functions that rate the presence of nuclear localisation signals.

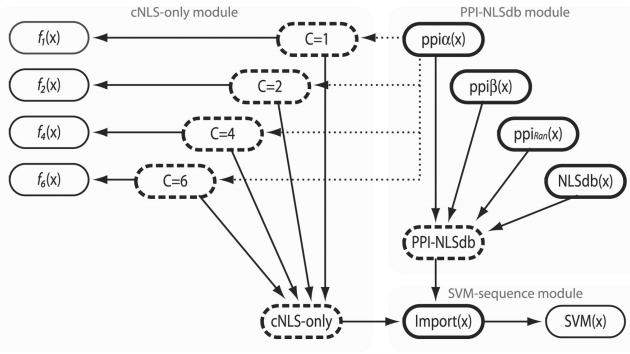


Fig. 1. A Bayesian Network for nuclear import. Ovals show random variables, solid arrows dependencies between them. Dashed nodes represent latent variables. Model v2 adds a dependency between the presence of cNLS on importin- α interaction (dashed arrows).

Kosugi et al. (2009) described six classes of motifs used in the importin- α import pathway. Of those six, one is only active in plants and one only present in very few proteins, so those were excluded. Mehdi et al. identified nuclear proteins containing any of the remaining four motifs and aligned those sequences centered around their motif. They created a position weight matrix (PWM) for each of the motif classes based on its respective alignment. The functions in the cNLS-only module use those PWMs to assign a score to the presence of a motif of each class in a given query sequence.

3.1.2 The PPI-NLSdb module The PPI-NLSdb module consists of four boolean random variables. Each of those shows, if a given query protein interacts with importin- α , importin- β or Ran, or contains an alternative nuclear localisation signal.

3.1.3 The SVM-sequence module The SVM-sequence module contains a support vector machine that rates the sequence similarity of the query protein with proteins that are known to localise in the nucleus.

The rating is based on the amino acid trimers in each sequence. In each sequence the number of occurrences of each possible trimer is counted and the sequence is then placed accordingly in a coordinate system with 8000 dimensions. The support vector machine then tries to separate the nuclear sequences from the non-nuclear sequences as well as possible.

3.2 Data sources

Mehdi et al. collected their data from various sources.

cNLS-only module: The cNLS motifs were described by Kosugi et al. (2009).

PPI-NLSdb module: The protein-protein interaction data was collected from the BioGRID protein-protein interactions data sets (Stark et al., 2006). As the coverage is very limited for mouse (only 9 interactions with importin- α , 32 with importin- β and 184 with Ran in mouse - 215, 375 and 132 in yeast respectively) indirect interactions, i.e. via one partner, were also included for the mouse predictions.

The NLS data was collected from NLSdb (Nair and Rost, 2003). This database contains motifs that have been experimentally verified

to play a role in nuclear import.

SVM-sequence module: For the training of the support vector machine a subset of the training sequences (NucProt, Fink et al. (2008) for mouse; yeast-GFP fusion data set, Huh et al. (2003) for yeast) were used. Those were excluded from the training for the network.

3.3 Validation

Mehdi and his colleagues performed six-fold cross-validation to evaluate their models (see Table 1 for results). They tested the single modules as well as two different versions of an integrative model. In version 1 the modules are connected only via the import node. In version 2 the hidden nodes for the different classes of nuclear localisation signals additionally depend on interaction with importin- α . Furthermore they compared the performance of their model to two state-of-the-art tools (NLStradamus and cNLS Mapper).

Both integrative models as well as the single modules perform significantly better than random (AUC greater than 0.5 and MCC greater than 0). The integrative models perform better than the single modules on both mouse and yeast data for both AUC and MCC. Version 1 and 2 both outperform cNLS Mapper and NLStradamus. Version 1 (without dependencies of the cNLS classes on importin- α interaction) performs slightly better than version 2 (AUC 0.84 vs 0.82, MCC 0.57 vs 0.52 on mouse, 0.80 vs 0.79 and 0.44 vs 0.42 on yeast respectively).

Table 1. Accuracy of predicting nuclear import with different models

Model	Mouse		Yeast	
	AUC	MCC	AUC	MCC
Model 1	0.84 \pm 0.02	0.57 \pm 0.02	0.80 \pm 0.01	0.44 \pm 0.01
Model 2	0.82 \pm 0.02	0.52 \pm 0.02	0.79 \pm 0.01	0.42 \pm 0.02
cNLS Mapper	0.66	0.29	0.61	0.24
NLStradamus	0.68	0.29	0.60	0.19
cNLS-only	0.71 \pm 0.01	0.31 \pm 0.01	0.70 \pm 0.01	0.24 \pm 0.01
PPI-NLSdb	0.62 \pm 0.01	0.16 \pm 0.01	0.60 \pm 0.01	0.16 \pm 0.01
SVM-sequence	0.78 \pm 0.01	0.51 \pm 0.01	0.76 \pm 0.01	0.37 \pm 0.01

Furthermore, Mehdi et al. tested their model on an independent data set (Hawkins et al., 2007) to evaluate the performance when trained on the entire available training data (see Table 2). Overlapping proteins between the training data and the independent data were removed from the training data, so the results are not due to duplicate proteins in the data sets. As an additional comparison, the prediction tool Nucleo, for which the data set had been created, was tested. Mehdi et al.'s model performed better than its contestants on the independent data set (except for a tie with Nucleo on the yeast proteins in the independent data set).

To ensure that the accuracy of their prediction was not caused by homologies within the data, the model was also tested on the independent data set with proteins with sequence similarity greater than 30% removed. As can be seen in Table 3 the quality of the prediction is affected by that restriction of the data, but only within a reasonable scale (MCC 0.56 to 0.50 for mouse and 0.32 to 0.41 for yeast respectively).

Table 2. Accuracy of prediction of nuclear import for proteins in an independent data set (Hawkins et al., 2007)

Model	Accuracy (MCC)		
	Mouse	Yeast	All species
Combined model	0.56	0.32	0.39
Nucleo	0.24	0.32	0.38
cNLS Mapper	0.41	0.26	0.27
NLStradamus	0.37	0.13	0.25

Table 3. Accuracy of nuclear import prediction for proteins with sequence similarity >30% removed.

Model	Accuracy(MCC)	
	Mouse	Yeast
Combined model	0.50	0.41
cNLS Mapper	0.28	0.26
NLStradamus	0.29	0.19

3.4 Discussion

In their paper Mehdi et al. present a novel method to predict nuclear import. It makes extensive use of the knowledge about the importin- α pathway, but also includes alternative nuclear localisation signals and sequence similarity to account for alternative or unknown localisation signals and pathways. They show the reliability of their prediction via cross-validation and on an independent data set. They also show that their method outperforms other state-of-the-art tools (cNLS Mapper, NLStradamus and Nucleo). Additionally their model provides mechanistic reasons for its decisions, so it can also be used to predict the presence of classical nuclear localisation signals and interaction with importin- α for a query protein.

4 SORTING THE NUCLEAR PROTEOME

Nuclear import is only the first highly regulated process that determines the function of the nucleus. A second regulatory step is the association of an imported protein with the right compartments at the right time.

The subnuclear compartments are highly dynamic complexes of protein, and potentially DNA and RNA. They do not need to be present in every cell or cell type and at any point of the cell cycle. Bauer et al. present a model that tries to predict the association of a given protein with subnuclear compartments.

But first, the predicted compartments shall be introduced briefly:

Nucleolus: The nucleolus is the biggest and best understood compartment. After mitosis one to four nucleoli form around clusters of ribosomal genes and mainly support transcription of rRNA and ribosomal biogenesis. The nucleolus is also known to play a role in cell-cycle control and stress response. Large-scale mass spectrometry experiments were able to find more than 700 proteins to associate with the nucleolus.

Perinucleolar compartment: The perinucleolar compartment mostly appears on the nucleolar surface in transformed and cancer cells. It is a very dynamic compartment that is present from late telophase

until the beginning of mitosis. Its main components are proteins involved in RNA metabolism and Polymerase III transcribed RNA. The perinucleolar compartment seems to be linked to an unknown DNA locus.

Promyelocytic leukaemia body: The Promyelocytic Leukaemia body host various processes including anti-viral response, DNA repair, apoptosis, gene regulation and tumor suppression. They are formed around the Promyelocytic Leukaemia (PML) protein and many proteins associated with the PML body are known to contain SUMOylation motifs and/or interaction sites. More than 75 proteins have been shown to interact with the PML body.

Nuclear speckle: The nuclear speckle is thought to play a role in pre-mRNA processing. It is a transit-zone for RNA binding proteins, so RNA binding motifs can be found in many proteins associated with the nuclear speckle.

Cajal body: The Cajal bodies seem to be a compartment where proteins involved in various nuclear processes concentrate and assemble to enhance their functional efficiency. Their core member is the protein coilin. It is involved in recruiting small ribonucleoproteins. The cajal bodies are very sensitive to changes in transcription.

Chromatin: Chromatin is the packaged form of DNA. It therefore consists of many DNA binding proteins and is responsible for granting access to the DNA itself.

Nuclear pore complex: The nuclear pore complexes (NPC) reside in the nuclear membrane and regulate the translocation of proteins from the cytoplasm into the nucleus and backwards. They are, compared to other subnuclear compartments, very static and present in all cells containing a nucleus.

Nuclear lamina: The nuclear lamina form a protein scaffold that stabilises the structure of the nucleus. They are also thought to play a role in transcriptional repression by remodeling the chromatin and binding to DNA.

4.1 Model

Unfortunately there is no precise mechanistic understanding of how subnuclear compartments are assembled, disassembled and regulated. Bauer et al. therefore rely on fairly generic information. They build a Bayesian Network that integrates protein-protein interaction data, domain and motif knowledge and sequence similarity with proteins known to associate with certain nuclear compartments (see Fig. 2).

Bauer et al. created one model for every compartment. They obtained knowledge about key protein members or scaffold proteins from the literature. If they were not able to identify four proteins, they looked at the correlation between the member proteins of a compartment and chose the ones with the highest correlation.

They also correlated domains and sequence motifs with membership in each compartment and integrated the four highest scoring ones into the respective network.

Finally, a support vector machine rates the sequence similarity of a query protein with the known members of a compartment.

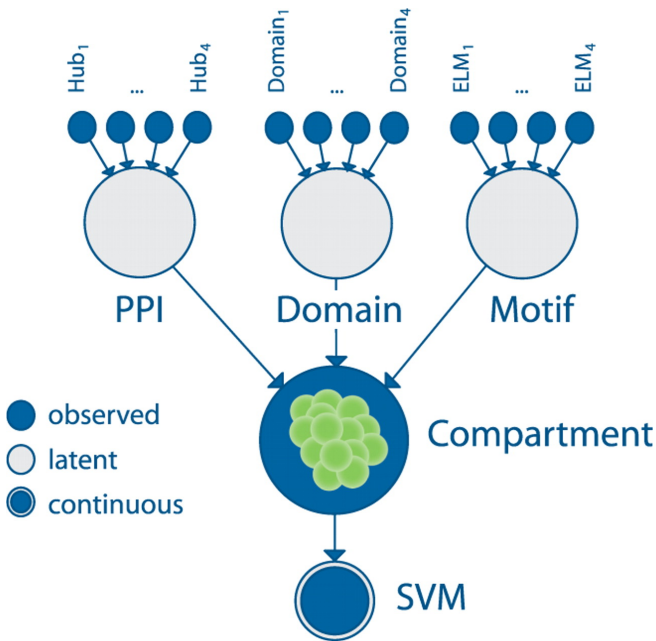


Fig. 2. A model for subnuclear compartment association

4.2 Data sources

The data was collected from many different sources. Protein data was obtained the mouse nuclear proteome (Fink et al., 2008), the Nuclear Protein Database (Dellaire et al., 2003), NOPdb (Leung et al., 2006), and other small data sets in the literature.

3567 proteins of the proteins were annotated as nuclear, 1286 of those with subnuclear compartment information.

Sequence and protein-protein interaction data was collected from Uniprot and BioGrid (Breitkreutz et al., 2008). InterPro and InterProScan (Hunter et al., 2009) were employed to find protein domains. Sequence motifs and post-translational modification sites were obtained from the Eukaryotic Linear Motif (ELM) resource (Gould et al., 2010).

4.3 Validation

Bauer et al. used the above training data to evaluate their model. Due to the lack of an independent data set and comparable tools, they were not able to validate their model on other data or rank the performance of their approach with regard to other techniques.

4.3.1 Cross-validation First of all they performed five-fold cross-validation on their data set (see Tab. 4 for results). The prediction for all compartments is substantially better than random ($AUC > 0.5$) with the worst predictions for the Cajal body and the Nucleolus with an AUC of 0.60 each. The best predicted compartments are the perinucleolar and the nuclear pore with an AUC of 0.80 and 0.79 respectively.

They also provide an AUC50 measure, that is the AUC measured up until 50 false positive predictions. As can also be seen in Table 4 the AUC drops strongly for all predicted compartments. The strongest drop occurs for Chromatin (0.71 to 0.17) and the PML

Table 4. Accuracy for prediction on proteins with known compartment association (five-fold cross-validation).

Compartment	Proteins	AUC50(SD)	AUC(SD)
Cajal body	51	0.22(0.02)	0.60(0.03)
Chromatin	323	0.17(0.02)	0.71(0.01)
Nuclear lamina	77	0.17(0.04)	0.70(0.01)
Nuclear pore	51	0.41(0.07)	0.79(0.05)
Nuclear specke	404	0.24(0.01)	0.71(0.01)
Nucleolus	596	0.14(0.01)	0.60(0.01)
Perinucleolar	24	0.41(0.09)	0.80(0.05)
PML body	91	0.23(0.06)	0.77(0.03)
Mean (compartment)		0.25	0.71

Table 5. Predicted association with subnuclear compartment for 2281 unannotated nuclear mouse proteins.

Compartment	Additional proteins	Probability threshold	FDR at threshold
Cajal body	23	0.24	0.76
Chromatin	509	0.43	0.52
Nuclear lamina	17	0.38	0.68
Nuclear pore	12	0.31	0.43
Nuclear specke	229	0.41	0.45
Nucleolus	1266	0.44	0.47
Perinucleolar	1	0.29	0.58
PML body	96	0.34	0.64

body (0.77 to 0.23). The Cajal body (0.60 to 0.22) and the nuclear pore (0.79 to 0.41) are affected the least.

This data shows that the overall prediction contains a lot of false positives and should therefore be looked at with caution.

4.3.2 Prediction of novel compartment association Bauer et al. then predicted compartment association for the unannotated proteins in their data set (2281 in total). The results can be seen in Table 5. The table shows the probability threshold above which proteins were assumed to be associated with a compartment if they were not assigned a false discovery rate (FDR; the estimated probability of a positive prediction being false) higher than the FDR threshold. As can be seen, compartment association was partly predicted even for very low probabilities and high false discovery rates (with the cajal body having the lowest probability/ highest FDR at 0.24/0.76 - three out of four proteins with such a low probability have to be assumed to be false although predicted as positive!).

This again hints towards a lot of false positive predictions.

To build trust in the top predictions of their model Bauer et al. looked at each compartment and its strongest prediction (see Table 6). As they used a Bayesian Network for their prediction they were able to trace back what information caused the high probability for compartment association. Then they turned to the literature and tried to find support for the predicted proteins or at least the reasons causing the prediction.

They found that the association of SSF1 with the nucleolus (predicted with a probability of 1.0) had previously been shown by Kim and Hirsch (1998). The same holds true for NSD1 and

Table 6. Top unannotated protein with predicted association for each subnuclear compartment.

Protein	Compartment	Probability	Est. FDR
ZN593(Q9DB42)	Nucleolus	1.00(0.00)	0.00
NFAT5(Q9WV30)	PML body	0.94(0.00)	0.00
IRF9(Q61179)	Cajal body	0.94(0.01)	0.00
PHF12(Q5SPL2)	Chromatin	0.92(0.00)	0.00
RUXF(P62307)	Nuclear speckle	0.92(0.00)	0.12
SMG1(Q8BKX6)	Nuclear pore	0.90(0.00)	0.20
ZFHX3(Q61329)	Nuclear lamina	0.78(0.01)	0.80
MINT(Q62504)	Perinucleolar	0.40(0.01)	0.62

HRX and their association with the chromatin (each predicted with a probability of 0.93, shown by Berdasco et al., 2009 and Guenther et al., 2005).

For all other proteins, except the perinucleolar one, they were able to provide properties of the respective proteins that made the association with its compartment reasonable.

Based on this information, the top predictions of the model seem to be correct, although it is not clear up until which probability and FDR threshold they can be trusted.

4.4 Discussion

Although the overall predictions are likely to contain a lot of false positives, the authors were able to show that the top predictions may well be reliable. Therefore it seems that annotations of subnuclear compartment associations are not complete. However, due to the lack of overall quality in the prediction, it is not possible to say how incomplete it is.

Therefore further experiments have to be conducted to enhance our understanding of subnuclear compartments – how they are regulated, when and how they assemble and disassemble and how they carry out their functions specifically. The predictive annotation by Bauer et al. can serve as a guide for those further experiments, if the considered proteins are predicted with high confidence to associate with the respective compartment and if that prediction is founded on reasonable biological causes.

REFERENCES

- [1] Denis C. Bauer, Kai Willadsen, Fabian A. Buske, Kim-Anh Lê Cao, Timothy L. Bailey, Graham Dellaire, and Mikael Bodn. Sorting the nuclear proteome. *Bioinformatics*, 27(13):i7–i14, 2011.
- [2] Ahmed M. Mehdi, Muhammad Shoaib B. Sehgal, Bostjan Kobe, Timothy L. Bailey, and Mikael Bodn. A probabilistic model of nuclear import of proteins. *Bioinformatics*, 27(9):1239–1246, 2011.